Environmental Toxicology and Chemistry—Volume 42, Number 6—pp. 1386–1400, 2023
Received: 12 January 2023 | Revised: 24 March 2023 | Accepted: 24 March 2023

Development of Multiple Linear Regression Models for Predicting Chronic Iron Toxicity to Aquatic Organisms

Kevin V. Brix,[a,b,*] Lucinda Tear,[c] David K. DeForest,[c] and William J. Adams[d]

[a]EcoTox, Miami, Florida, USA
[b]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida, USA
[c]Windward Environmental, Seattle, Washington, USA
[d]Red Cap Consulting, Lake Point, Utah, USA

Abstract: We developed multiple linear regression (MLR) models for predicting iron (Fe) toxicity to aquatic organisms for use in deriving site-specific water quality guidelines (WQGs). The effects of dissolved organic carbon (DOC), hardness, and pH on Fe toxicity to three representative taxa (*Ceriodaphnia dubia*, *Pimephales promelas*, and *Raphidocelis subcapitata*) were evaluated. Both DOC and pH were identified as toxicity-modifying factors (TMFs) for *P. promelas* and *R. subcapitata*, whereas only DOC was a TMF for *C. dubia*. The MLR models based on effective concentration 10% and 20% values were developed and performed reasonably well, with adjusted R^2 of 0.68–0.89 across all species and statistical endpoints. Differences among species in the MLR models precluded development of a pooled model. Instead, the species-specific models were assumed to be representative of invertebrates, fish, and algae and were applied accordingly to normalize toxicity data. The species sensitivity distribution (SSD) included standard laboratory toxicity data and effects data from mesocosm experiments on aquatic insects, with aquatic insects being the predominant taxa in the lowest quartile of the SSD. Using the European Union approach for deriving WQGs, application of MLR models to this SSD resulted in WQGs ranging from 114 to 765 $\mu g\ l^{-1}$ Fe across the TMF conditions evaluated (DOC: 0.5–10 $mg\ l^{-1}$; pH: 6.0–8.4), with slightly higher WQGs (199–910 $\mu g\ l^{-1}$) derived using the US Environmental Protection Agency (USEPA) methodology. An important uncertainty in these derivations is the applicability of the *C. dubia* MLR model (no pH parameter) to aquatic insects, and understanding the pH sensitivity of aquatic insects to Fe toxicity is a research priority. An Excel-based tool for calculating Fe WQGs using both European Union and USEPA approaches across a range of TMF conditions is provided. *Environ Toxicol Chem* 2023;42:1386–1400. © 2023 The Authors. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Iron; Bioavailability; Multiple linear regression; Water quality guidelines

INTRODUCTION

Iron (Fe) is an abundant metal, comprising 5% of the Earth's crust, and is essential for life in oxygen-rich environments, where it is used for respiration in animals and photosynthesis in plants (Bury et al., 2012). Ferrous iron (Fe II) is more soluble and generally more bioavailable than ferric iron (Fe III; Bury et al., 2012; Lofts et al., 2008). Most concerns with Fe toxicity in freshwater environments are downstream from iron ore mining activities and acid mine drainage (Bury et al., 2012). Existing freshwater quality guidelines for Fe generally range between 300 $\mu g\ l^{-1}$ (Canadian Council of Resources and Environment Ministers, 1987) and 1000 $\mu g\ l^{-1}$ total recoverable Fe (US Environmental Protection Agency [USEPA], 1976). These existing guidelines are based on limited data sets and so have high uncertainty. For example, the USEPA's chronic water quality criterion is based largely on observations at an Fe-impacted stream in Colorado (USA), where trout and other fish species were not observed at Fe concentrations >1000 $\mu g\ l^{-1}$ (USEPA, 1976).

As one approach for addressing uncertainty in Fe criteria, Linton et al. (2007) evaluated field bioassessment data using quantile regression to establish total Fe benchmarks and suggested a benchmark of 1250 $\mu g\ l^{-1}$ although they noted some insect taxa were sensitive to Fe at concentrations as low as 210 $\mu g\ l^{-1}$. Similarly, under the Water Framework Directive, the

This article includes online-only Supporting Information.

* Address correspondence to KevinBrix@EcoTox.onmicrosoft.com
Published online 29 March 2023 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5623

UK Technical Advisory Group has proposed a new Fe quality standard in freshwaters of 730 μg l^{-1} total Fe (Peters et al., 2012). As with the USEPA Fe criterion, it is also primarily based on field evidence that showed no decline in fish, macrophyte, and diatom communities, but impacts above this concentration on macroinvertebrate communities at pH > 7.0.

The mechanisms of Fe toxicity to aquatic organisms are not well studied, particularly in chronic exposures. The prevailing hypothesis is that Fe(III) can coat the gills of aquatic animals, increasing oxygen diffusion distance and leading to respiratory impairment. This hypothesis is partially supported by experiments on Atlantic salmon (*Salmo salar*) exposed to 500 μg l^{-1} at pH 6.3 where Fe(II) is the predominant species. Under these condition, little Fe accumulation at the gill and no mortality was observed, but when pH was increased to 6.7, Fe(III) hydroxide species formed and precipitated from solution onto the gill, and increasing mortality was observed (Teien et al., 2008). In an earlier study, Peuranen et al. (1994) exposed brown trout (*Salmo trutta*) to 2000 μg l^{-1} Fe(II) at pH 5–6 and observed fusion of gill lamellae leading to increased oxygen diffusion distances and reduced oxygen consumption. This finding suggests that at somewhat higher concentrations than used by Teien et al. (2008), structural effects on the gill may also contribute to overall respiratory stress. Ionoregulatory disturbances (changes in plasma Na and Ca concentrations) have also been observed in several fish species exposed to Fe, but they have been associated with concurrent increases in cortisol and lactate, suggesting that this is not a direct interaction with ion transporters but rather is associated with a generalized stress response to impaired respiration (Bury et al., 2012; Lappivaara et al., 1999; Peuranen et al., 1994). The primary mechanisms of Fe(III) toxicity to algae appears to be removal of essential nutrients, particularly phosphate. Arbildua et al. (2017) conducted a series of experiments on the green alga *Raphidocelis supcaptitata* demonstrating that toxicity correlated well with removal of phosphate from solution by Fe coprecipitation and that toxicity could be eliminated by addition of phosphorous to the solution after the initial precipitation event.

It has been known for several decades that Fe toxicity in laboratory studies can be greater than that observed in field studies. In addition, it is known that ambient waters with varying alkalinity, pH, hardness, temperature, and ligands that influence the valence state and solubility of Fe may all influence Fe toxicity in a given water (USEPA, 1976). However, to our knowledge, research on how varying key water chemistry parameters influence Fe toxicity to a taxonomically diverse set of freshwater organisms has not been systematically tested. Most studies have focused on how pH influences Fe speciation and the differential toxicity of Fe(II) and Fe(III) (Gerhardt, 1992; Teien et al., 2008), but results are not always consistent between taxa. For example, Gerhardt (1992) found that the mayfly *Leptophlebia marginata* was more sensitive to Fe(II) at pH 4.5 than Fe(III) at pH 7, contrary to the observations of Teien et al. (2008) on Atlantic salmon (*S. salar*). A few studies have explored the effects of other toxicity-modifying factors (TMFs) on Fe toxicity. For example, Peuranen et al. (1994) evaluated the influence of humic acid (15 mg l^{-1}) and found that it ameliorated the effects of Fe on ion regulation and oxygen uptake. Lappivaara et al. (1999) likewise found that the physiological effects (cortisol levels, ethoxyresorufin-O-deethylase activity, plasma ion composition) of Fe were negligible after exposure to humic acid‑rich runoff water from a peat production area.

In a companion article in this issue, the influence of dissolved organic carbon (DOC), hardness, and pH on chronic Fe (added as Fe(III)) toxicity to an alga (*Raphidocelis subcapitata*), an invertebrate (*Ceriodaphnia dubia*), and a fish (*Pimephales promelas*) are characterized (Cardwell et al., 2023 [companion paper]). The objective of our study was to use these data to develop empirical bioavailability models that predict Fe toxicity as a function of these TMFs. To accomplish this, we used a multiple linear regression (MLR) model approach consistent with those previously described for Al, Cu, and Zn (Brix et al., 2017, 2020, 2021; DeForest et al., 2018, 2020, 2023).

MATERIALS AND METHODS

Toxicity data sources

The methods and results of toxicity testing conducted with *R. subcapitata*, *C. dubia*, and *P. promelas* are provided in a companion paper (Cardwell et al., 2023; Supporting Information, Tables S1–S3). These were the only data identified for development of species-specific MLRs. Additional data were identified in the literature for development of species sensitivity distributions (SSDs) and are summarized in the Supporting Information, Table S4. These studies and those described in Cardwell et al. (2023) generally met USEPA guidelines (Stephan et al., 1985) for acceptable chronic toxicity tests for use in ambient water quality criteria (AWQC) derivation, with the exception of the 7-day fathead minnow studies, which do not meet the USEPA guideline requirements for a chronic toxicity study for fish. Although inclusion of the 7-day data for fathead minnow likely underestimates the chronic sensitivity of this species, we concluded that the information gained by including data for this species across a wide range of water chemistry conditions outweighed the uncertainty introduced by using data from subchronic experiments. This same rationale was used by DeForest et al. (2018, 2020) in developing an MLR-based WQC for Al, which was subsequently adopted by the USEPA (2018).

To understand how the toxicity data described in Cardwell et al. (2023) are distributed as a function of pH, DOC, and hardness, we compared them with a previously described natural waters data set from the United States (Brix et al., 2020). This comparison was conducted using previously described methods (Brix et al., 2021) in which a principal component analysis (PCA) of the *z*-scores of log(DOC), log(Hardness), and pH from the United States natural waters data set was conducted in R (*prcomp*()) to reduce the data set to two dimensions. The PCA model was then used to predict axis scores for the log-transformed toxicity data in R (*predict*()), and a scatterplot of the scores for each species toxicity data set was overlain on a scatterplot for the natural waters data set from the United States to provide a visual comparison.

MLR analysis

Stepwise MLR analysis was conducted to evaluate whether chronic Fe toxicity to *R. subcapitata*, *C. dubia*, and *P. promelas* could be modeled as a linear function of DOC, hardness, and pH. The methods for this analysis followed those previously described for other MLR models for Al, Cu, and Zn (Brix et al., 2020, 2021; DeForest et al., 2020, 2023).

Briefly, stepwise MLR analyses were conducted using an initial model including ln(DOC), ln(hardness [Hard]), and pH as independent variables and ln(EC*X*) as the dependent variable. Initial models assumed that ln(DOC), ln(Hard), and pH had no interactions and used the general form:

$$\ln(ECX) = b_0 + b_1 \times \ln(DOC) + b_2 \times \ln(Hard) + b_3 \times pH \quad (1)$$

where EC*X* = effect concentration (EC)10 or EC20 ($\mu g\ l^{-1}$); b_0 = the intercept; b_1, b_2, and b_3 are the slopes for ln(DOC), ln(Hard), and pH, respectively; and DOC and hardness are in units of mg l^{-1}.

A subsequent analysis allowed for interactions between the main independent variables and included all two-way interactions:

$$\begin{aligned}\ln(ECX) = {} & b_0 + b_1 \times \ln(DOC) + b_2 \times \ln(Hard) + b_3 \times pH \\ & + b_4 \times \ln(DOC) \times pH \\ & + b_5 \times \ln(DOC) \times \ln(Hard) \\ & + b_6 \times \ln(Hard) \times pH\end{aligned} \quad (2)$$

where EC*X*, b_0, b_1, b_2, and b_3 are defined as above, and b_4, b_5, and b_6 are the slopes for ln(DOC) × pH, ln(DOC) × ln(Hard), and ln(Hard) × pH, respectively.

Models based on both the EC10 and EC20 were developed to support application under different regulatory frameworks. Best-fitting models were identified using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC; Burnham & Anderson, 2004). The stepwise model with the fewest terms, as determined using AIC or BIC, was considered the best-fitting model. Stepwise regressions were conducted in R using the stepAIC() function from the MASS library (e.g., for the BIC model: stepAIC [model, direction = c("both"), k = 1n(*n*)]; Venables & Ripley, 2002). Variance inflation factors were calculated in R using the vif() function in the usdm library (Naimi et al., 2014).

Model evaluation

We evaluated species-specific and pooled models using several different metrics, as previously described in MLR models for Al, Cu, and Zn (Brix et al., 2021; DeForest et al., 2020). Initial assessments of model performance used standard diagnostic tools for evaluating regressions including adjusted R^2, predicted R^2, residuals analysis, and variance inflation factors (VIFs; Burnham & Anderson, 2004; Harrell, 2015; Zuur et al., 2010). Adjusted R^2 was calculated to assess the proportion of the variability in Fe toxicity values that could be explained by the model, adjusting for the number of terms in the model and the number of discrete observations. Predicted R^2 was calculated to assess the ability of the model to predict new data, substituting the Predictive Residual Sums of Squares (PRESS) for the Total Residual Sums of Squares (TSS) in the R^2 calculation. The predicted R^2 will always be at least slightly lower than the adjusted R^2; large decreases in the predicted R^2 relative to the adjusted R^2 indicate that the model may be overparameterized or strongly influenced by one or more tests in the data set. Although no a priori criteria exist to determine how large a difference is meaningful, we generally consider a predicted $R^2 > 0.10$ lower than the adjusted R^2 to merit further evaluation and a reduction >0.20 to indicate significant problems with the model.

Following the model evaluation guidelines outlined in Garman et al. (2020), we also conducted four additional performance evaluations. First we developed 1:1 plots of observed versus predicted EC*X* values, which include diagonal lines denoting 1:1 and factor of ±2 agreement. The ±2 agreement lines were selected as bounding parameters because they are consistent with interlaboratory variability associated with toxicity test methods (Moore et al., 2000) and so provide a reasonable approximation of the expected level of accuracy. We also evaluated model residuals of observed versus predicted values as a function of observed toxicity, DOC, pH, and hardness to identify whether there were systematic biases in the model predictions. We then developed cumulative probability distributions of observed/predicted EC*X* values and factor of agreement plots to expand on information in the 1:1 plots included in the MLR models along with a null model against which to compare MLR (Garman et al., 2020).

Model validation

There are a number of potential approaches to model validation including autovalidation, independent validation, and cross-species validation (Garman et al., 2020) as well as randomization and simulation techniques. Given the limited number of species ($n = 3$) and number of toxicity tests/species ($n = 18–27$), we selected a randomization approach called *k*-fold cross-validation to evaluate model performance (Hastie et al., 2009) that we also used to validate MLRs for Zn (DeForest et al., 2023). In this approach, the data set for each species is randomized and then divided into *k* subsets ("folds"). Each cross-validation consists of fitting a "training" MLR model to predict toxicity for the data in *k*-1 of the subsets and then using the training model to predict toxicity in the unused "test" subset. In each run, a different subset is used as the test subset so that each of the original subsets eventually serves as a test subset. Given the size of each of the species-specific MLR data sets and the general observation that *k*'s of 5 or 10 tend to produce useful results in many situations, $k = 5$ was selected to preserve a minimum of three samples in each fold. Using $k = 5$, each subset consists of 20% of the data; training models are fit to four subsets (80% of the data) at a time and tested on the remaining one subset (20% of the data).

The AIC/BIC selected model for each species and statistical endpoint was used as the training model in each cross-validation run. Model coefficients and fit parameters (predicted R^2 and Root Mean Square Error [RMSE]; Equation 3) were calculated for each of the five training models along with the RMSE of each test subset's fit to the training model.

$$\mathrm{RMSE} = \sqrt{\frac{SS_{\mathrm{residuals}}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x})^2}{n}} \tag{3}$$

where $SS_{\mathrm{residuals}}$ is the sum of squared residuals, $x_i = \ln(\text{observed ECX})$, and $\hat{x} = \text{predicted } \ln(\text{ECX})$.

The mean, standard deviation, and coefficient of variation (CV) of each TMF coefficient, significance level (p value), and predicted R^2 were calculated across the five training models to provide insights into how model coefficients and fit varied based on the data used. Variation in significance values for each TMF in the five training sets was used to evaluate confidence in the final selected TMFs for each model. Training and test RMSE were compared to provide insight about how well a training model based on 80% of the data might fit the other 20% of the data. Finally, for each training model, water quality guidelines (WQGs) were derived using the methods described in the following section as a function of each model TMF to provide insight into the implications of how the variability characterized in the validation exercise would influence the WQG.

Derivation of chronic Fe water quality guidelines

The final component of our study was to evaluate how the chronic Fe toxicity data for *R. subcapitata*, *C. dubia*, and *P. promelas*, and the resulting MLR-based models for predicting chronic Fe toxicity as a function of TMFs, could be used to derive a WQG for Fe. Chronic Fe(III) toxicity data were compiled from other studies, with minimum requirements including that the Fe concentrations were measured in the toxicity test and that the endpoint was directly related to effects on survival, growth, development, or reproduction. In addition to using single-species laboratory studies, we also considered the use of effects data generated from mesocosm studies. This allowed us to increase the representation of insect taxa in the SSD and genus sensitivity distribution (GSD), which are typically significantly under-represented relative to natural aquatic communities due to the limited number of standardized laboratory test protocols available for this group (Brix et al., 2005; Mebane et al., 2020). Details regarding data selection and additional statistical analyses conducted to support data used in the SSD are provided in Supporting Information.

As discussed in the *Species-specific Fe MLR models* section, we found significant differences between models for each of the three species evaluated, such that it was not possible to pool MLR models across species. Although the *R. subcapitata*, *C. dubia*, and *P. promelas* Fe MLR models have not been validated for other species, we assumed that the *C. dubia* and *P. promelas* models would be relevant to other invertebrate and fish species, respectively. (No chronic Fe toxicity data were identified for other algal or plant species.) We recognize there is uncertainty associated with applying these models to other taxa, but this approach is consistent with the cross-species extrapolation of biotic ligand models (BLMs) for environmental quality standard development in the European Union (Schlekat et al., 2010; Van Sprang et al., 2009).

The empirical EC10s and EC20s were normalized to a water chemistry of interest using Equation (4).

$$\begin{aligned}\ln(ECX_{\mathrm{norm}}) = {} & \ln(ECX_{\mathrm{meas}}) - b_1[\ln(DOC_{\mathrm{meas}}) - \ln(DOC_{\mathrm{site}})] \\ & - b_2[pH_{\mathrm{meas}} - pH_{\mathrm{site}}] - b_3[(\ln Hard_{\mathrm{meas}}) \\ & - (\ln Hard_{\mathrm{site}})]\end{aligned} \tag{4}$$

In Equation (4), ECX_{norm} is the Fe EC10/20 normalized to the site water chemistry, ECX_{meas} is the empirically measured EC10/20, b_1–b_3 are slopes for the TMFs, DOC_{meas} is the DOC measured in the toxicity test, and DOC_{site} is the DOC in the site water of interest, with the same terminology used for pH and hardness (Hard). This is the generalized equation with the actual parameters retained adjusted on a species-specific basis to those TMFs identified as significant in the MLR analysis.

Once the toxicity data were normalized to a water chemistry of interest, SSDs and GSDs were developed based on the species and genus mean values following European Union and USEPA guidance (Stephan et al., 1985; Suter, 2002; van Straalen & Van Leeuwen, 2002) to estimate WQGs based on EC10s (European Union approach) and EC20s (USEPA approach).

The final output from this analysis was the development of a hazardous concentration to 5% of the species (HC5) calculator that derives potential WQGs for Fe as a function of TMFs consistent with European Union and USEPA guidelines (see the HC5 Calculator in the Supporting Information). The HC5 calculator is Excel-based and is similar to the calculator developed for Al (DeForest et al., 2018, 2020).

RESULTS

Distribution of toxicity data compared with natural waters

The combinations of TMFs used in chronic toxicity testing for *C. dubia*, *P. promelas*, and *R. subcapitata* were relatively well distributed compared with the distribution found in natural waters of the United States (Figure 1). Testing for all three species was skewed somewhat toward waters with relatively low hardness and pH compared with the distribution in natural waters. However, this was a necessary consequence of Fe having relatively low toxicity and issues with Fe precipitation in waters with high hardness and pH (Cardwell et al., 2023) rather than an intentional avoidance of naturally occurring conditions in the toxicity testing study design.

Species-specific Fe MLR models

The stepwise MLR analyses resulted in development of models with and without interactions for each species (Table 1 and Supporting Information, Table S4). In general, AIC and BIC selected the same models, except for the *C. dubia* models with interactions, where AIC retained more interaction terms than BIC. When TMF interactions were allowed, there were only three instances when interaction terms were retained in the model, and in these cases, predicted R^2 increased by no more than 0.05 relative to models without interactions (Table 1). Given these results, subsequent model development and evaluation focused on models without interactions.

Species-specific models retained the same parameters and had similar slopes for a given species when predicting both EC10s and EC20s (Table 1 and Supporting Information, Table S4). Variance inflation factors were low (<1.2) for all models, indicating low correlation between TMFs within the toxicity data sets (Supporting Information, Table S5). For simplicity, we primarily discuss the EC10 models throughout the remainder of our study, but results for both statistical endpoints are provided in either the main manuscript or the Supporting Information. The equations for predicting *C. dubia*, *P. promelas*, and *R. subcapitata* Fe EC10s as a function of water chemistry are provided in Equations (5–7), respectively.

$$\text{Predicted } C.\ dubia \text{ Fe EC10} = \exp[7.577 + 0.600 \times \ln(\text{DOC})] \tag{5}$$

$$\text{Predicted } P.\ promelas \text{ Fe EC10} = \exp[2.176 + 1.102 \times \ln(\text{DOC}) + 0.787 \times \text{pH}] \tag{6}$$

$$\text{Predicted } R.\ subcapitata \text{ Fe EC10} = \exp[5.435 + 0.744 \times \ln(\text{DOC}) + 0.332 \times \text{pH}] \tag{7}$$

The *C. dubia* MLR model retained only DOC as a TMF and had the lowest performance metrics of the three species-specific models with an adjusted R^2 of 0.74 and predicted R^2 of 0.71 (Table 1 and Figure 2). The $RF_{x,2.0}$ of 81% is an

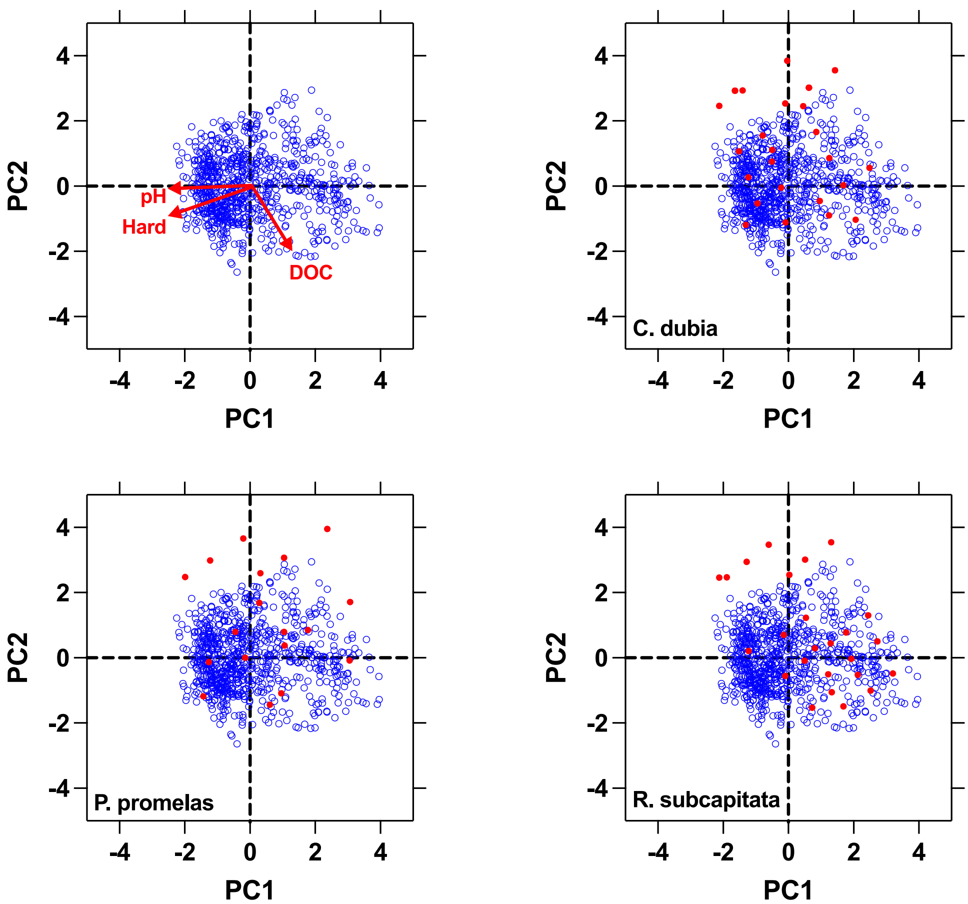

FIGURE 1: Biplots of axis scores for principal component (PC)1 (60% of variance) and PC2 (33% of variance) of principal component analysis of Z-scores of log(dissolved organic carbon [DOC]), log(Hardness), and pH. Plot of natural waters data set with red arrows showing eigenvectors for toxicity-modifying factors and plots of predicted axis scores of chronic toxicity data (red circles) compared with PC1 and PC2 scores of natural waters data set (light blue circles) for *Ceriodaphnia dubia*, *Pimephales promelas*, and *Raphidocelis subcapitata*.

TABLE 1: Summary of species-specific Fe multiple linear regression models selected by Bayesian Information Criterion

Species	Interactions	Intercept	ln(DOC)	ln(Hard)	pH	ln(DOC) × ln(Hard)	ln(DOC) × pH	ln(Hard) × pH	No.	Adj. R^2	Pred. R^2
EC10											
Ceriodaphnia dubia	No	7.577	0.600						27	0.74	0.71
	Yes	7.554	1.363	0.006		−0.175			27	0.78	0.70
Pimephales promelas	No	2.176	1.102		0.787				18	0.84	0.81
	Yes	2.176	1.102		0.787				18	0.84	0.81
Raphidocelis subcapitata	No	5.435	0.744		0.332				25	0.87	0.84
	Yes	5.435	0.744		0.332				25	0.87	0.84
EC20											
C. dubia	No	7.925	0.507						27	0.68	0.63
	Yes	13.777	1.330	−1.499	−0.828	−0.192		0.212	27	0.75	0.68
P. promelas	No	2.607	0.920		0.830				18	0.89	0.87
	Yes	2.607	0.920		0.830				18	0.89	0.87
R. subcapitata	No	6.914	0.541		0.173				25	0.87	0.84
	Yes	2.606	0.536	0.957	0.836			−0.147	25	0.89	0.85

Adj. = adjusted; DOC = dissolved organic carbon; EC10, 20 = effective concentration 10%, 20%; Pred. = predicted.

improvement over the null model $RF_{x,2.0}$ of 63% (Figure 3). The EC10 model residuals revealed no significant relationship with observed toxicity (Figure 4). In contrast, a significant relationship was found between observed toxicity and EC20 model residuals, indicating that the EC20 model tends to underpredict toxicity in waters with low toxicity and overpredict toxicity in waters with high toxicity (Supporting Information, Figure S3). However, no significant pattern was observed between residuals and hardness, DOC, or pH, suggesting the modeled TMFs are not strongly influencing the residual pattern as a function of toxicity (Supporting Information, Figure S3).

The MLR model for *P. promelas* retained both DOC and pH as TMFs with the DOC slope nearly twice as steep as observed for *C. dubia* and a pH slope nearly twice as steep as that of *R. subcapitata* (Table 1). The *P. promelas* MLR model performs well with an adjusted R^2 of 0.84 and predicted R^2 of 0.81 (Table 1 and Figure 2). The $RF_{x,2.0}$ of 78% provides a large improvement over the null model $RF_{x,2.0}$ of 11% (Figure 3). No significant patterns in residuals were observed for either the EC10 or EC20 models for *P. promelas* (Figure 5 and Supporting Information, Figure S4).

The MLR model for *R. subcapitata* also retained DOC and pH as TMFs and performed well with an adjusted R^2 of 0.87 and predicted R^2 of 0.84 (Table 1 and Figure 2). The $RF_{x,2.0}$ of 100% provides an improvement over the null model $RF_{x,2.0}$ of 48% (Figure 3). Although no statistically significant linear patterns were observed in model residuals, a clear inverted U-shaped pattern can be seen in the relationship between EC10 model residuals and observed toxicity (Figure 6). Although less distinct, a similar pattern is also observed in the relationship between residuals and DOC, suggesting that this TMF may be contributing to the pattern in residuals relative to observed toxicity (Figure 6). A similar pattern is not readily apparent in the relationship between EC20 model residuals and observed toxicity, but is still present when plotted as a function of DOC (Supporting Information, Figure S5).

Differences in model parameters and parameter slopes across species were considered large enough to preclude development of pooled species models.

Model validation

Results for the *k*-fold validation exercise were generally similar for EC10 and EC20 models. We focus on the results for EC10 models with similar information for EC20 models provided in the Supporting Information, Table S6.

We first evaluated model stability by characterizing consistency in model parameter selection and variability in parameter coefficients across the 5 folds compared with the model based on all data for each species. Parameter selection was generally consistent using AIC and BIC. For simplicity, we focus on EC10 results using BIC, but results using both endpoints and statistical methods can be found in the Supporting Information, Table S6. Dissolved organic carbon was consistently selected as a TMF in all training data sets for all three species. Similarly, pH was consistently selected as a TMF for *P. promelas* and *R. subcapitata*, but was selected in only one training model for *C. dubia*. Hardness was not selected as a TMF for any *C. dubia* training model, but was selected in one training model for both *P. promelas* and *R. subcapitata* (Supporting Information, Table S6).

The AIC/BIC selected models were also evaluated by comparing the adjusted R^2, predicted R^2, training model RMSE, and test model RMSE for each of the 5 folds to the selected MLR model for each species using all data (Table 2 and Supporting Information, Table S6). Variability in model coefficients for selected models differed across species, but patterns in variability were broadly similar between EC10 and EC20 models (Table 3 and Supporting Information, Table S6). For EC10 training models in which the same TMFs were selected as for the MLR based on the full data set, variability in DOC slopes was low (CV = 4%–9%). Variability in pH slopes was higher than DOC slopes for *P. promelas* and *R. subcapitata* with a CV of 18% for both species.

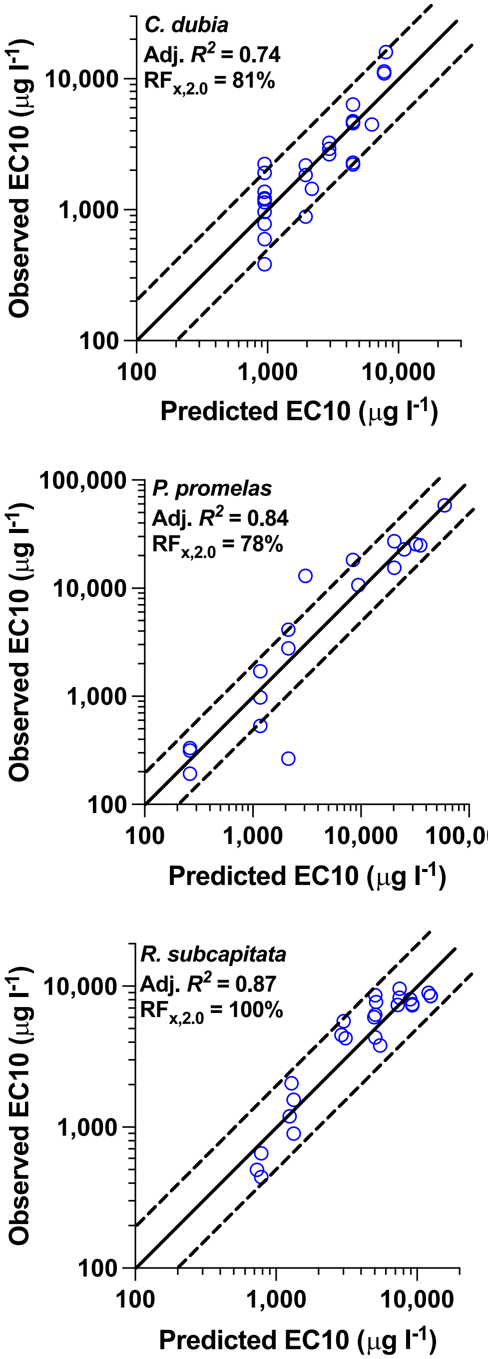

FIGURE 2: Comparative 1:1 plots of the chronic effective concentration, 10% (EC10) Fe multiple linear regression (MLR) models for *Ceriodaphnia dubia*, *Pimephales promelas*, and *Raphidocelis subcapitata*. Solid line is line of perfect agreement between predicted and observed EC10s. Dashed lines indicate ± a factor of 2.

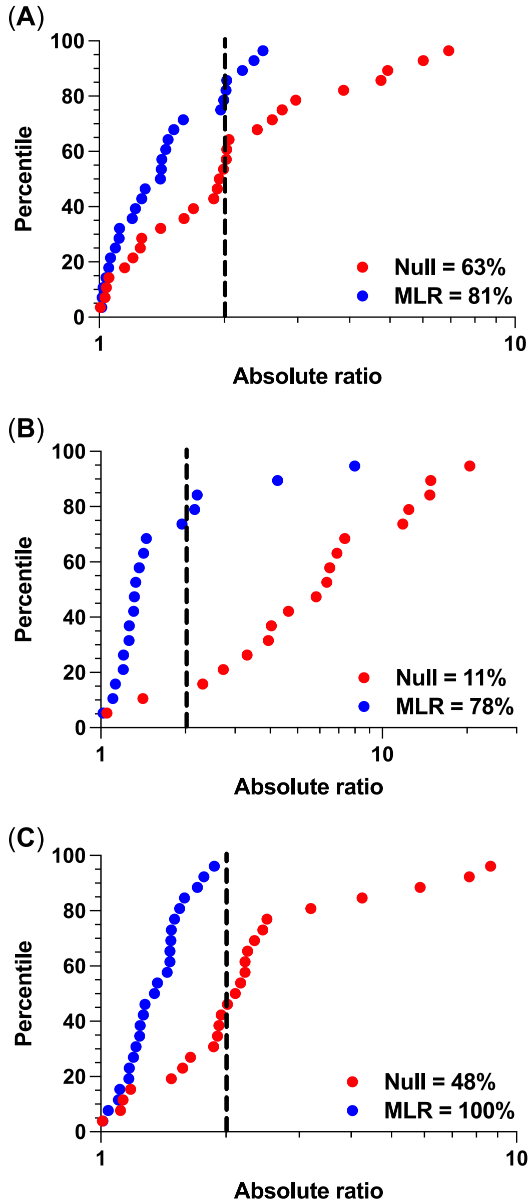

FIGURE 3: Cumulative density functions of predicted–observed ratios for null and multiple linear regression (MLR) effective concentration, 10% (EC10) models for (**A**) *Ceriodaphnia dubia*, (**B**) *Pimephales promelas*, and (**C**) *Raphidocelis subcapitata*. Absolute ratios calculated as $10^{|\log(\text{predicted/observed})|}$. Dashed lines indicate ± a factor of 2.

Coefficients of variation for model intercepts were very low for *C. dubia* (less than 1%), intermediate for *R. subcapitata* (8%), and high for *P. promelas* (47%).

Chronic Fe WQGs

The Fe toxicity data and MLR models generated from our study, after being augmented with additional chronic toxicity data available in the literature, were used to develop potential chronic freshwater quality guidelines for Fe. The SSD ($n = 28$) and GSD ($n = 26$) are nearly equivalent, with only the genus *Daphnia* having two species representatives; the algae *R. subcapitata* was also removed from the GSD, consistent with USEPA methodology (USEPA, 1985; Supporting Information, Table S7). The SSD is comprised of 1 amphibian, 5 fish, 1 algal, and 21 invertebrate species including data for 16 insect species that were primarily derived from a mesocosm study (Cadmus et al., 2018a, 2018b).

Differences in MLR models across taxa precluded use of a single equation for deriving a WQG. Instead, toxicity data for

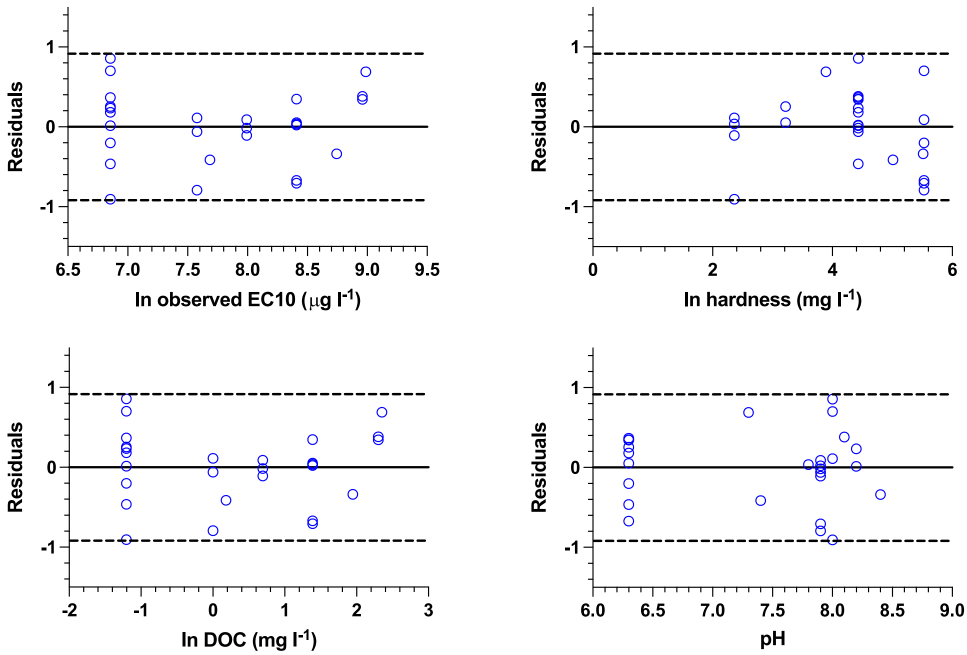

FIGURE 4: *Ceriodaphnia dubia* chronic effective concentration, 10% (EC10) Fe multiple linear regression (MLR) model residuals plotted as a function of observed toxicity, hardness, dissolved organic carbon (DOC), and pH. Dotted lines indicate ±2 standard deviations of residuals for all data used to evaluate the model.

each species were normalized to any specified water quality conditions using the representative MLR model for each taxonomic group (invertebrate, fish/amphibian, algae). A WQG calculator tool was developed in Excel that normalized all toxicity data and estimates an HC5 for any specified water quality conditions (Supporting Information).

For both the European Union-based SSD and the USEPA-based GSD, the most sensitive taxa are insects under most water quality conditions. However, under low pH and DOC conditions, mountain whitefish (*Prosipium williamsoni*) becomes the most sensitive taxa (Figure 7) primarily due to the relatively steep pH slope for the fish MLR, whereas the

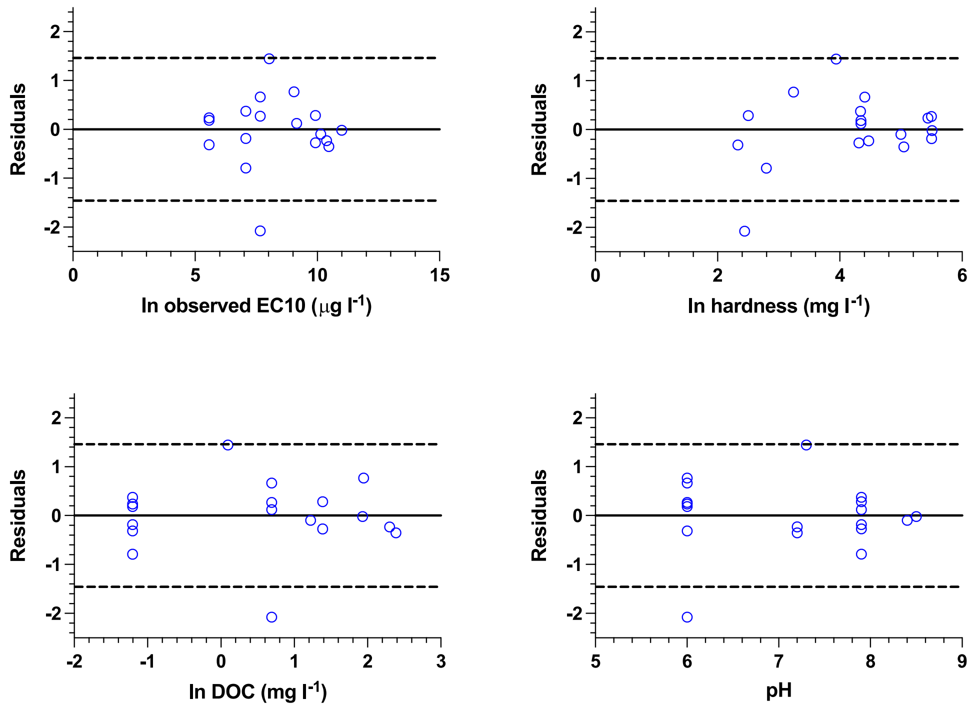

FIGURE 5: *Pimephales promelas* chronic effective concentration, 10% (EC10) Fe multiple linear regression (MLR) model residuals plotted as a function of observed toxicity, hardness, dissolved organic carbon (DOC), and pH. Dotted lines indicate ±2 standard deviations of residuals for all data used to evaluate the model.

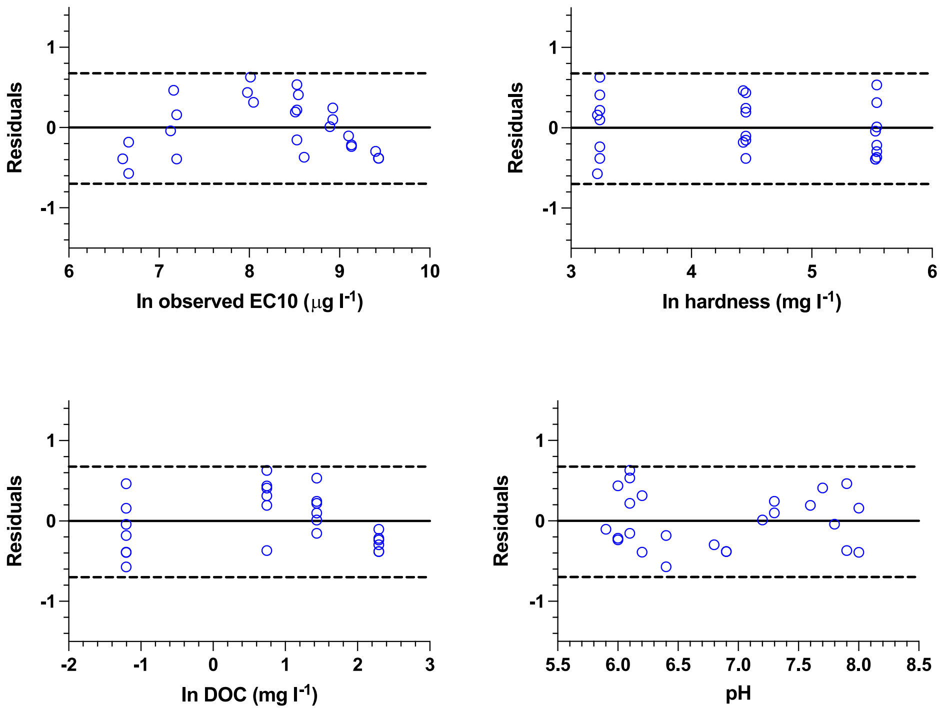

FIGURE 6: *Raphidocelis subcapitata* chronic effective concentration, 10% (EC10) Fe multiple linear regression model residuals plotted as a function of observed toxicity, hardness, dissolved organic carbon (DOC), and pH. Dotted lines indicate ±2 standard deviations of residuals for all data used to evaluate the model.

invertebrate MLR does not include pH as a TMF. The pH at which mountain whitefish becomes the most sensitive taxa varies from 6.9 (European Union SSD) to 6.3 (USEPA GSD) due to differences in pH slopes between the EC10 and EC20 MLR models. One consequence of invertebrates being the most sensitive taxa in the SSD under alkaline water quality conditions is that WQGs are a function of only DOC for the USEPA methodology. The same is largely true for the European Union methodology, although changes in fish and algae sensitivity as a function of pH have a modest influence on the estimated HC5 (Figure 8).

Uncertainty in HC5 estimates were evaluated as part of the *k*-fold validation procedure. Model parameters derived for

TABLE 2: Summary of *k*-fold validation results for 10% effect concentration Fe multiple linear regressions selected by Bayesian Information Criterion with k1–k5 representing the 5-folds evaluated

Model	Adj. R^2	Pred. R^2	Training RMSE	Test RMSE
Ceriodaphnia dubia				
k1	0.86	0.81	0.340	0.930
k2	0.66	0.60	0.456	0.461
k3	0.67	0.61	0.435	0.525
k4	0.77	0.74	0.448	0.482
k5	0.75	0.71	0.453	0.427
Average	0.74	0.70	0.426	0.565
Pimephales promelas				
k1	0.81	0.76	0.766	0.145
k2	0.91	0.89	0.498	1.266
k3	0.82	0.78	0.733	0.579
k4	0.80	0.74	0.701	0.830
k5	0.87	0.82	0.620	0.966
Average	0.84	0.80	0.664	0.757
Raphidocelis subcapitata				
k1	0.91	0.89	0.300	0.500
k2	0.85	0.80	0.346	0.329
k3	0.87	0.83	0.353	0.282
k4	0.88	0.85	0.282	0.523
k5	0.83	0.77	0.353	0.287
Average	0.87	0.83	0.327	0.384

Adj. = adjusted; Pred. = predicted; RMSE = root mean square error.

TABLE 3: Summary of *k*-fold validation parameters for 10% effect concentration Fe multiple linear regressions

Model	5-Fold training models			Species model
	Mean	Range	CV	
Ceriodaphnia dubia				
Intercept	7.599	7.53–7.64	0.006	7.577
DOC	0.605	0.52–0.64	0.090	0.600
Pimephales promelas				
Intercept	2.220	1.49–3.72	0.468	2.176
DOC	1.091	1.03–1.19	0.061	1.102
pH	0.783	0.58–0.88	0.177	0.787
Raphidocelis subcapitata				
Intercept	5.489	4.91–5.98	0.082	5.435
DOC	0.744	0.71–0.79	0.048	0.744
pH	0.322	0.26–0.39	0.176	0.332

Mean, range, and coefficient of variation of each model parameter across 5 folds compared with the species model. All model selections are based on BIC. Only training models that retained the same parameters as the final model were evaluated.
BIC = Bayesian Inforrmation Criterion; CV = coefficient of variation; DOC = dissolved organic carbon.

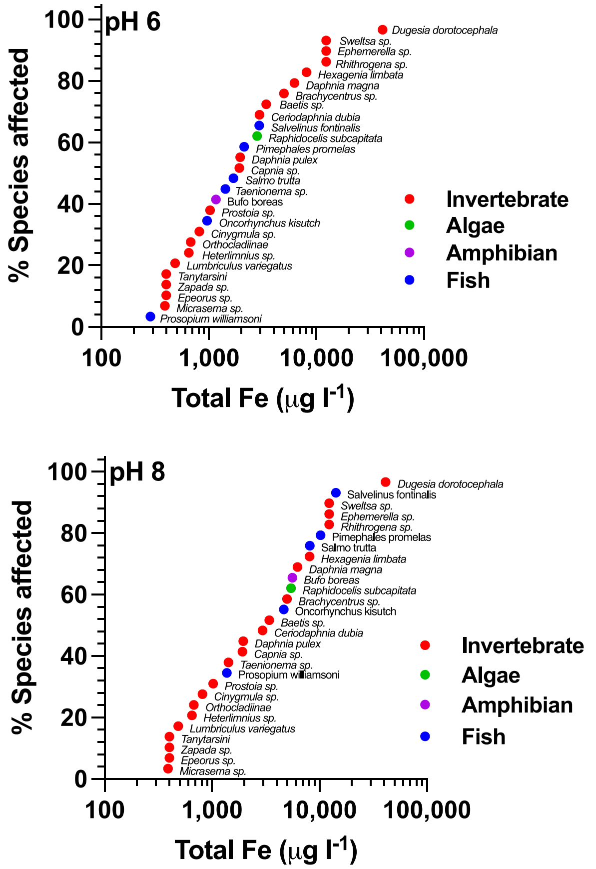

FIGURE 7: Species sensitivity distributions for Fe based on the European Union methodology at pH 6 and 8 at dissolved organic carbon = 2 mg l^{-1}, hardness = 100 mg l^{-}. Note large differences in relative sensitivity of fish at pH 6 versus pH 8.

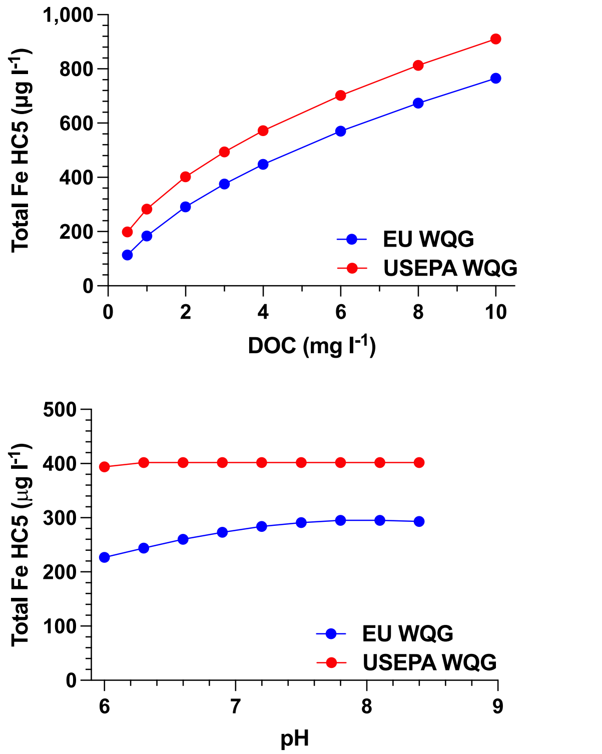

FIGURE 8: Comparison of chronic hazardous concentrations for 5% of the species (HC5s) for Fe based on European Union (EU) and USEPA methodologies as a function of dissolved organic carbon (DOC) and pH. The DOC-dependent HC5s were calculated at pH = 7.5, hardness = 100 mg l^{-1}. The pH-dependent HC5s were calculated at DOC = 2 mg l^{-1}, hardness = 100 mg l^{-1}. WQG = water quality guideline.

each of the training data sets were used to estimate HC5s as a function of DOC and pH. Results from this analysis indicate that HC5 estimates are relatively insensitive to variation in the toxicity data sets used to derive the MLR models (Figure 9 and Spporting Information, Figure S6). For example, using the European Union WQG approach, CVs in HC5s ranged from 0.8% to 11.7% across a series of DOC concentrations from 0.5 to 10 mg l^{-1}. For pH, where one of the *k*-fold training models for *C. dubia* identified pH as a significant TMF whereas the remaining did not (as was the case for the *C. dubia* model using the full data set), CVs in HC5s were still low, ranging from 1.3% to 11.9% across pH conditions from 6.0 to 8.4 (Figure 9).

DISCUSSION

MLR model development and performance

Our study aimed to develop chronic Fe MLR models for use in deriving site-specific WQGs. The MLR models were developed for three phylogenetically diverse taxa relying on relatively small toxicity data sets that were generated specifically for the purpose of developing bioavailability models (Cardwell et al., 2023). Models for all three species indicate that DOC is an important TMF for Fe. Although there were differences in DOC slopes across taxa, the range in DOC slopes for Fe was generally comparable to those observed for Cu and Al (Brix et al., 2021; DeForest et al., 2020), and higher than those for Ni and Zn (Croteau et al., 2021; DeForest et al., 2023). Considering the affinity of DOC for Fe, the importance of DOC in mitigating bioavailability is not surprising (Lofts et al., 2008).

Hardness was not a statistically significant term in MLR models for the three species evaluated (Table 1). Iron is the only metal evaluated to date using an MLR approach for which hardness is not a significant TMF. Iron speciation is dominated by Fe(III) over the range of water chemistry conditions evaluated, and so it may not be surprising from a mechanistic perspective that the divalent cations (Ca and Mg) that comprise hardness do not have a significant effect on the bioavailability or toxicity of a trivalent cation. However, hardness has been identified as an important TMF for Al, which is also generally present as a trivalent cation over the same pH range (DeForest et al., 2020), suggesting that valence state may not be an explanatory mechanism for Fe toxicity not being hardness dependent.

As expected, a strong effect of pH on Fe toxicity was observed for *P. promelas* and a more modest effect for *R. subcapitata*. Surprisingly though, no pH effect on Fe toxicity was observed for *C. dubia*. One weakness of the MLR approach is the potential for a relatively robust TMF effect to not be detected if data characterizing that effect are limited relative to

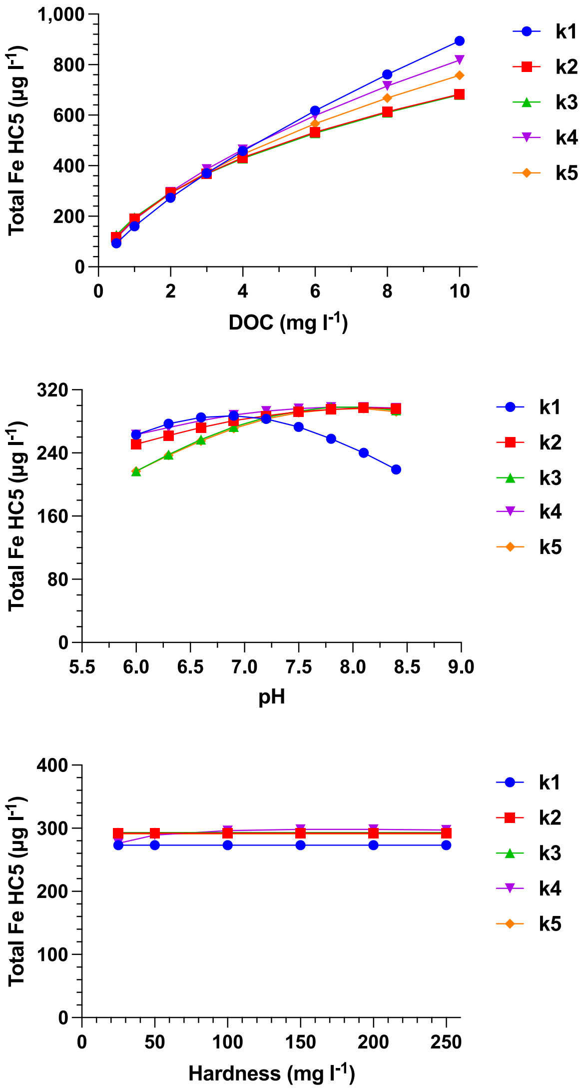

FIGURE 9: Comparison of *k*-fold multiple linear regression models for deriving chronic hazardous concentrations for 5% of the species (HC5s) for Fe using the European Union method as a function of dissolved organic carbon (DOC) and pH. The DOC-dependent HC5s were calculated at pH = 7.5, hardness = 100 mg l^{-1}. The pH-dependent HC5s were calculated at DOC = 2 mg l^{-1}, hardness = 100 mg l^{-1}. The hardness-dependent HC5s were calculated at DOC = 2 mg l^{-1}, pH = 7.5.

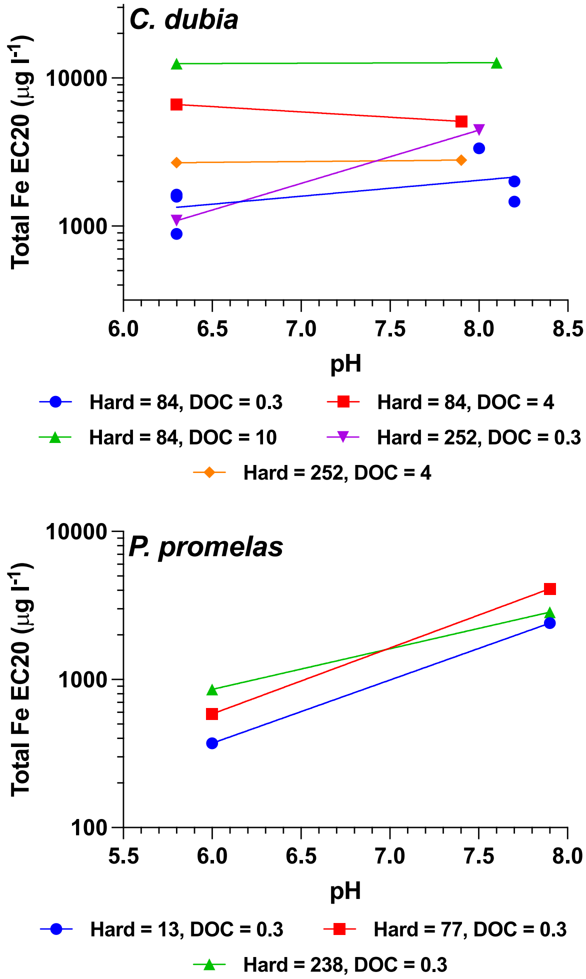

FIGURE 10: Univariate evaluation of the effect of pH on Fe toxicity to *Ceriodaphnia dubia* and *Pimephales promelas*. In some cases, small variations in hardness occurred between experiments for *P. promelas*, and the average hardness is represented in the plot. EC20 = effective concentration, 20%; DOC = dissolved organic carbon.

the overall data set. To address this issue, we typically plot subsets of the data where only one TMF is varied (i.e., univariate data) to evaluate whether univariate data are consistent with the MLR model. Evaluation of available data suggests that the MLR model for *C. dubia* is likely correct in excluding pH as a TMF (Figure 9). Most of the univariate slopes are near zero, with the exception of experiments conducted at a hardness of 252 mg l^{-1} and DOC of 0.3 mg l^{-1}, where the slope is 0.108 based on ln-transformed EC20 data. In contrast, univariate experiments with *P. promelas* consistently exhibit a strong pH effect, with slopes for ln-transformed EC20s ranging from 0.632 to 1.03, consistent with the MLR model slope of 0.830 (Figure 10 and Table 1).

The lack of a consistent pH response across taxa prevented development of pooled MLR models for WQG derivation. Instead, the *C. dubia* model was applied to all invertebrates, the *P. promelas* model to all fish, and the *R. subcapitata* model to all plants. This is the same approach previously used to develop a site-specific WQG for Al (DeForest et al., 2018, 2020) and is consistent with the approach commonly used in Europe for BLM-based WQG (European Commission, 2011). However, making the assumption that MLRs for a given species are applicable to large taxonomic groups (e.g., invertebrates) is uncertain. This assumption has been supported in some cases, but not in others, in previous studies with other metals where bioavailability models have been derived for multiple species (Brix et al., 2021; Croteau et al., 2021; DeForest et al., 2023; Schlekat et al., 2010).

Using a variety of metrics, the overall performance of the species-specific Fe MLR models was found to be generally comparable or better than MLR models for other metals (Brix et al., 2021; Croteau et al., 2021; DeForest et al., 2020). The *P.*

promelas and *R. subcapitata* MLRs generally performed better than the *C. dubia* MLRs. To some extent, this may be explained by the higher $RF_{x,2.0}$ for the *C. dubia* raw data (Figure 3). In general, when the raw data $RF_{x,2.0}$ is relatively high, as is the case for the Fe *C. dubia* data set (63%), then a relatively small fraction of the data set is sufficiently influenced by TMFs to exceed the approximate ± factor of 2 variability inherent to the test method. Consequently, much of the variance in the data set is the result of test method variability and is not readily predicted by the MLR. Of course, this type of scenario can also be viewed as indicating that metal toxicity for this species is not strongly influenced by the TMF(s) under consideration, as is the case for *C. dubia*, for which only DOC was found to be a significant TMF.

For model validation, we used the *k*-fold validation approach we first described in developing Zn MLRs (DeForest et al., 2023). This approach is particularly useful for metals like Fe, for which the toxicity data sets are relatively small and exclusion of data for independent validation results in even smaller training and validation data sets with limited statistical power and ability to identify statistically significant relationships. The resampling used in the *k*-fold approach for Fe provides an empirical evaluation of the variance in the TMFs retained by AIC/BIC for each MLR model using the full data set. For example, DOC was highly statistically significant ($p < 0.00005$) in the model using the full data set and was also consistently selected as a TMF across all folds for all species in the cross-validation training models (Supporting Informaion, Table S6). Similarly, pH was significant for both *P. promelas* and *R. subcapitata* in the models based on the full data sets ($p < 0.005$) and was consistently selected as a TMF in training models of the cross-validation. In contrast, while pH was not selected for *C. dubia* and hardness not selected for *P. promelas* and *R. subcapitata*, these parameters were selected in at least one of the training models, suggesting that their exclusion from models using the full data set is less certain (Supporting Informaion, Table S6).

Variability in model parameters was relatively low across folds, with CVs generally 20% or less across EC10 and EC20 models (Table 3 and Supporting Information, Table S6). The one exception was for the *P. promelas* EC10 model where the CV for the model intercept was 47%. This variability was largely driven by one training model (*k*2), which had a substantially higher model intercept and low DOC slope compared with the other models, as well as inclusion of hardness as a TMF in two of the models.

The information generated in the *k*-fold validation regarding uncertainty in the models can also be extended to provide information on uncertainty in corresponding WQGs derived by applying these models. To evaluate this, we used each of the *k*-fold training models to derive HC5s across a range of pH and DOC conditions (Figure 9 and Supporting Information, Table S6). Variability in the *k*-fold MLR models (Table 3) generally translated to comparable variability in HC5s (as measured by CVs) across a range of DOC concentrations. The HC5s were generally more variable at low pH than at neutral or alkaline conditions, with the exception of one training set in which a negative pH term was retained for the *C. dubia* model. Despite the inclusion of hardness in several of the training models for *P. promelas* and *R. subcapitata*, HC5s varied little as a function of hardness. Overall, our analysis suggests that estimated HC5s as a function of TMFs are relatively insensitive to uncertainty in MLR model parameter estimates.

Iron water quality guidelines

In developing the SSD and GSD for Fe, we departed from the more typical approach of relying on only single-species laboratory toxicity studies and included data from a stream mesocosm study (Cadmus et al., 2018a, 2018b). This is not a novel approach and has been applied to other metals including Cd, Ni, and Zn when sufficient concentration–response data have been developed in mesocosm experiments (Mebane et al., 2020). Our inclusion of mesocosm data is motivated by the under-representation of insects in SSDs even though this group is often the dominant (by species and abundance) class of macrofauna present in many aquatic systems (Brix et al., 2005; Buchwalter et al., 2017; Mebane et al., 2020). This under-representation is primarily driven by the limited number of standard laboratory test methods for aquatic insects (Brix et al., 2005, 2011).

In addition to increasing biological diversity in the SSD, inclusion of mesocosm data also provides data under more realistic exposure scenarios that capture potential contributions of dietary metal exposure to effects, effects of metal exposure on diet availability and quality, and competitive interactions between taxa (Brix et al., 2011; Clements, 2004; Mebane et al., 2020). Although these characteristics of mesocosm studies are desirable for deriving WQGs, they also complicate interpretation of mesocosm data and consequently applicability of the data on a site-specific basis. In addition, there is some uncertainty regarding the repeatability of mesocosm studies relative to standard laboratory toxicity tests. The repeatability of mesocosm data has not been well studied, but the available information suggests they may be more variable than standard laboratory tests, with individual taxa sensitivity varying in response to starting abundance and likely other factors (Mebane et al., 2020). Overall though, we conclude that these uncertainties are outweighed by the added value of including mesocosm data in the SSD.

An alternative, or additional, approach to including mesocosm data in the SSD is consideration or use of field data in setting WQG using techniques such as quantile regression (Schmidt et al., 2012). This approach has been applied in several studies for Fe as well as other metals (Crane et al., 2007; Linton et al., 2007; Peters et al., 2012, 2014). Specific to Fe, Linton et al. (2007) evaluated macroinvertebrate bioassessment data from West Virginia (USA). They estimated EC20s using quantile regression for eight insect families, with Leptophlebiidae (Ephemeroptera) being the most sensitive at a total Fe EC20 of 210 $\mu g\,l^{-1}$, and suggested this value would be appropriately protective of sensitive aquatic taxa. Crane et al. (2007) evaluated bioassessment data from England and Wales and used higher level metrics (including Ephemeroptera, Plecoptera, and Trichoptera, as well as Ephemeroptera) to measure taxa richness rather than evaluating individual species. Using estimated EC10s for these endpoints, they concluded that an appropriate total Fe quality standard would be in the range of 43–250 μg l^{-1}.

Peters et al. (2012) expanded on the analysis by Crane et al. (2007) by evaluating individual insect families. Interestingly, five of the seven taxa identified in the Linton et al. (2007) study as being sensitive to Fe were also identified as sensitive (lower quartile of sensitivity) by Peters et al. (2012), suggesting congruence in relative sensitivity of insect taxa using data sets from different continents. Peters et al. (2012) recommended a total Fe quality standard of 730 μg l^{-1}, not taking into consideration the effects of TMFs on Fe bioavailability in the waters used to derive the quality standard. To address this issue, they used a subset ($n = 8$) of the *C. dubia* data published in Cardwell et al. (2023) to derive an MLR model that included hardness and DOC as TMFs to normalize results. Results from this analysis indicated that the quality standard would be reduced by a factor of 2.5 (292 μg l^{-1}) under high bioavailability conditions, which were defined as a hardness of 15 mg l^{-1} and DOC of 1.1 mg l^{-1}. Because pH was not reported for their high bioavailability normalizing conditions, there is some uncertainty regarding what the estimated WQG would be using the approach described in our analysis. However, assuming hardness and pH are relatively well correlated, a pH of 6.5 is a reasonable approximation of the conditions in a water with a hardness of 15 mg l^{-1}. Under these conditions, the resulting WQGs would be 163 and 297 μg l^{-1} Fe, using the European Union and USEPA methodologies, respectively.

Overall, these previous studies using field-based approaches derive WQGs that are generally similar to the guidelines we derived using an SSD approach based on single species and mesocosm toxicity data. Under high bioavailability conditions (<1 mg l^{-1} DOC, pH < 7), both field-based approaches and our approach estimate an Fe WQG of approximately 100–200 μg l^{-1}. Although these values are somewhat lower than would be derived using only single-species laboratory data, their congruence suggests that these lower values are more appropriate for protecting sensitive insect taxa.

Uncertainties and limitations

Development of an Fe WQG that considers bioavailability would provide a marked improvement over existing guidelines that do not consider TMFs. This is clearly evidenced by the improved predictability of toxicity compared with a null model for all three species for which MLR models were developed (Figure 3). However, there are some uncertainties and limitations of our approach that should be considered with respect to implementation and future research.

An important limitation of the existing MLR models, and therefore a limitation in their application, is the lack of data below pH 6. The potential for Fe toxicity in acidic drainages such as those associated with acid mine drainage is a concern (Schmidt et al., 2002). Iron toxicity in acidic drainages is associated with significant changes in Fe speciation at pH < 6, with Fe^{2+} becoming increasingly prevalent. Consequently, despite these types of environments being one of the more likely environments in which Fe toxicity to aquatic organisms might be observed, we do not recommend that the current MLR models be used to estimate Fe toxicity or WQG beyond the environmental conditions used to derive the models.

A further uncertainty related to pH is the lack of a pH response in the *C. dubia* MLR model. This finding is somewhat surprising given the strong response of Fe speciation to changing pH and the relatively robust response of the *P. promelas* MLR, and to a lesser extent the *R. subcapitata* MLR, to pH. It is worth noting that the "low pH" water (pH 6.3) evaluated for MLR development with *C. dubia* is at the lower end of the tolerance range for this species. This is evidenced by the generally lower control reproductive output at pH 6.3 compared with high pH conditions (Cardwell et al., 2023). However, this apparent pH stress does not appear to interact with Fe given the lack of a pH response for *C. dubia* (Figure 10). Although this lack of response appears to be well supported, the application of this model to all invertebrates is uncertain. This is important because the lower quartile of the Fe SSD is dominated by aquatic insects, most of which were tested at pH 7.2 (Figure 7). If these taxa respond to pH similar to that for fish, then WQGs derived using the current framework may be underprotective under more acidic conditions (e.g., pH 6) and overprotective under more alkaline conditions (e.g., pH 8). There is a clear need for testing the sensitivity of aquatic insects to Fe under different pH conditions.

CONCLUSIONS

Our objective was to use the toxicity data described in Cardwell et al. (2023) to develop MLR models for use in deriving site-specific chronic WQGs for Fe. Models using MLR were successfully developed for all three taxa evaluated (*C. dubia*, *P. promelas*, and *R. subcapitata*). Dissolved organic carbon was consistently identified as a TMF across all three taxa, whereas hardness was consistently not a TMF. The pH response was variable across taxa, with *P. promelas* being the most sensitive, whereas pH was not a TMF for *C. dubia*. As a result, a pooled MLR model was not developed, and species-specific models were applied by taxonomic groups (invertebrate, fish, algae) to derive site-specific WQGs.

We developed SSDs using both European Union and USEPA approaches based on a combination of standard laboratory toxicity data and mesocosm data from a study on aquatic insects. The mesocosm data proved to be influential, because aquatic insects dominate the lower quartile of the SSD. Integrating the SSD and MLR models using the European Union approach results in WQGs ranging from 114 to 765 μg l^{-1} Fe across the range of TMF conditions for which the model is applicable (Figure 8). An important uncertainty in these derivations is the applicability of the *C. dubia* MLR model (no pH parameter) to aquatic insects, and understanding the pH sensitivity of aquatic insects to Fe toxicity is a research priority. An Excel-based tool is provided for calculation of Fe WQGs under water quality conditions for which the model is applicable (Supporting Information).

Supporting Information—The Supporting Information is available on the Wiley Online Library at https://doi.org/10.1002/etc.5623.

Acknowledgments—This study was funded by the Iron Platform and Rio Tinto.

Disclaimer—The authors declare no conflict of interest.

Author Contributions Statement—**Kevin V. Brix**: Conceptualization; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing—original draft; Writing—review & editing. **Lucinda Tear**: Data curation; Formal analysis; Investigation; Writing—review & editing. **David K. DeForest**: Data curation; Formal analysis; Investigation; Writing—review & editing. **William J. Adams**: Conceptualization; Funding acquisition; Project administration; Supervision; Writing—review & editing.

Data Availability Statement—All data used in the present study are provided as Supporting Information.

REFERENCES

Arbildua, J. J., Villavicencio, G., Urrestarazu, P., Opazo, M., Brix, K. V., Adams, W. J., & Rodriguez, P. H. (2017). Effects of Fe (III) on *Pseudokirchneriella subcapitata* at circumneutral pH in standard laboratory tests is explained by nutrient sequestration. *Environmental Toxicology and Chemistry*, *36*, 952–958.

Brix, K. V., DeForest, D. K., & Adams, W. J. (2011). The sensitivity of aquatic insects to divalent metals: A comparative analysis of laboratory and field data. *Science of the Total Environment*, *409*, 4187–4197.

Brix, K. V., DeForest, D. K., Burger, M., & Adams, W. J. (2005). Assessing the relative sensitivity of aquatic organisms to divalent metals and their representation in toxicity data sets compared to natural aquatic communities. *Human and Ecological Risk Assessment*, *11*, 1139–1156.

Brix, K. V., DeForest, D. K., Tear, L., Grosell, M., & Adams, W. J. (2017). Use of multiple linear regression models for setting water quality criteria for copper: A complementary approach to the biotic ligand model. *Environmental Science & Technology*, *51*, 5182–5192.

Brix, K. V., DeForest, D. K., Tear, L., Peijnenburg, W. J. G. M., Peters, A., Middleton, E. T., & Erickson, R. J. (2020). Development of empirical bioavailability models for metals. *Environmental Toxicology and Chemistry*, *39*, 85–100.

Brix, K. V., Tear, L., Santore, R. C., Croteau, K., & DeForest, D. K. (2021). Comparative performance of multiple linear regression and biotic ligand models for estimating the bioavailability of copper in freshwater. *Environmental Toxicology and Chemistry*, *40*, 1649–1661.

Buchwalter, D. B., Clements, W. H., & Luoma, S. N. (2017). Modernizing water quality criteria in the United States: A need to expand the definition of acceptable data. *Environmental Toxicology and Chemistry*, *36*, 285–291.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–303.

Bury, N. R., Boyle, D., & Cooper, C. A. (2012). Iron. In C. M. Wood, A. P. Farrell, & C. J. Brauner (Eds.), *Homeostasis and toxicology of essential metals* (pp. 201–251). Elsevier.

Cadmus, P., Brinkman, S. F., & May, M. K. (2018a). Chronic toxicity of ferric iron for North American aquatic organisms: Derivation of a chronic water quality criterion using single species and mescosm data. *Archives of Environmental Contamination and Toxicology*, *74*, 605–618.

Cadmus, P., Guasch, H., Herdrich, A. T., Bonet, B., Urrea, G., & Clements, W. H. (2018b). Structural and functional reponses of periphyton and macroinvertebrate communities to ferric Fe, Cu, and Zn in stream mesocosms. *Environmental Toxicology and Chemistry*, *37*, 1320–1329.

Cardwell, A. S., Rodriguez, P. H., Stubblefield, W. A., DeForest, D. K., & Adams, W. J. (2023). Chronic toxicity of iron to aquatic organisms under variable pH, hardness, and dissolved organic carbon conditions. *Environmental Toxicology and Chemistry*, *42*, 1371–1385. https://doi.org/10.1002/etc.5627

Canadian Council of Resources and Environment Ministers. (1987). Canadian water quality guidelines.

Clements, W. H. (2004). Small-scale experiments support causal relationships between metal contamination and macroinvertebrate community responses. *Ecological Applications*, *14*, 954–967.

Crane, M., Kwok, K. W. H., Wells, C., Whitehouse, P., & Lui, G. C. S. (2007). Use of field data to support European water framework directive quality standards for dissolved metals. *Environmental Science & Technology*, *41*, 5014–5021.

Croteau, K., Ryan, A. C., Santore, R. C., DeForest, D. K., Schlekat, C. E., Middleton, E. T., & Garman, E. (2021). Comparison of multiple linear regression and biotic ligand models to predict the toxicity of nickel to aquatic freshwater organisms. *Environmental Toxicology and Chemistry*, *40*, 2189–2205.

DeForest, D. K., Brix, K. V., Tear, L. M., & Adams, W. J. (2018). Multiple linear regression (MLR) models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environmental Toxicology and Chemistry*, *37*, 80–90.

DeForest, D. K., Brix, K. V., Tear, L. M., Cardwell, A. S., Stubblefield, W. A., Nordheim, E., & Adams, W. J. (2020). Updated multiple linear regression (MLR) models for predicting chronic aluminum toxicity to freshwater aquatic organisms and development water quality guidelines. *Environmental Toxicology and Chemistry*, *39*, 1724–1736.

DeForest, D. K., Ryan, A. C., Tear, L. M., & Brix, K. V. (2023). Comparison of multiple linear regression and biotic ligand models for predicting acute and chronic zinc toxicity to freshwater organisms. *Environmental Toxicology and Chemistry*, *42*, 393–413.

European Commission. (2011). Common implementation strategy for the water framework directive (2000/60/EC). Guidance document no. 27. Technical guidance for deriving environmental quality standards, European communities.

Garman, E. R., Meyer, J. S., Bergeron, C. M., Blewett, T. A., Clements, W. H., Elias, M. C., Farley, K. J., Gissi, F., & Ryan, A. C. (2020). Validation of bioavailability-based toxicity models for metals. *Environmental Toxicology and Chemistry*, *39*, 101–117.

Gerhardt, A. (1992). Effects of subacute doses of iron (Fe) on *Leptophlebia marginata* (Insecta: Ephemeroptera). *Freshwater Biology*, *27*, 79–84.

Harrell, F. (2015). *Regression modeling stategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Lappivaara, J., Kiviniemi, A., & Oikari, A. (1999). Bioaccumulation and subchronic physiological effects of waterborne iron overload on whitefish exposed in humic and nonhumic water. *Archives of Environmental Contamination and Toxicology*, *37*, 196–204.

Linton, T. K., Pacheco, M. A. W., McIntyre, D. O., Clements, W. H., & Goodrich-Mahoney, J. (2007). Development of bioassessment-based benchmarks for iron. *Environmental Toxicology and Chemistry*, *26*, 1291–1298.

Lofts, S., Tipping, E., & Hamilton-Taylor, J. (2008). The chemical speciation of Fe^{3+} in freshwaters. *Aquatic Geochemistry*, *14*, 337–358.

Mebane, C. A., Schmidt, T. S., Miller, J. L., & Balistrieri, L. S. (2020). Bioaccumulation and toxicity of cadmium, copper, nickel and zinc and their mixtures to aquatic insect communities. *Environmental Toxicology and Chemistry*, *39*, 812–833.

Moore, D. W., Warren-Hicks, W., Parkhurst, B. R., Teed, R. S., Baird, R. B., Berger, R., Denton, D. L., & Pletl, J. J. (2000). Intra- and interlaboratory variabiilty in reference toxicity tests: Implications for whole effluent toxicity testing programs. *Environmental Toxicology and Chemistry*, *19*, 105–112.

Naimi, B., Hamm, N. A. S., Goren, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography*, *37*, 191–203.

Peters, A., Merrington, G., Simpson, P., & Crane, M. 2012. *Proposed quality standards for iron in freshwaters based on field evidence. Water Framework Directive—United Kingdom Technical Advisory Group*. SNIFFER/Environmental Agency, Edinburgh, United Kingdom.

Peters, A., Simpson, P., Merrington, G., Schlekat, C. E., & Rogevich-Garman, E. (2014). Assessment of the effects of nickel on benthic macroinvertebrates in the field. *Environmental Science and Pollution Research*, *21*, 193–204.

Peuranen, S., Vuorinen, P. J., Vuorinen, M., & Hollender, A. (1994). The effects of iron, humic acids and low pH on the gills and physiology of brown trout (*Salmo trutta*). *Annales Zoologici Fennici*, *31*, 389–396.

Schlekat, C. E., Van Genderen, E. J., De Schamphelaere, K. A. C., Antunes, P. M. C., Rogevich, E. C., & Stubblefield, W. A. (2010). Cross-species extrapolation of chronic nickel Biotic Ligand Models. *Science of the Total Environment*, *408*, 6148–6157.

Schmidt, T. S., Clements, W. H., & Cade, B. S. (2012). Estimating risks to aquatic life using quantile regression. *Freshwater Science*, *31*, 709–723.

Schmidt, T. S., Soucek, D. J., & Cherry, D. S. (2002). Integrative assessment of benthic macroinvertebrate community impairment from metal-contaminated waters in tributaries of the upper Powell River, Virginia, USA. *Environmental Toxicology and Chemistry*, *21*, 2233–2241.

Stephan, C. E., Mount, D. I., Hansen, D. J., Gentile, J. H., Chapman, G. A., & Brungs, W. A. (1985). *Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses*. U.S. Environmental Protection Agency, Environmental Research Laboratory, Duluth, MN.

Suter, G. W. (2002). North American history of species sensitivity distributions. In L. Posthuma, G. W. Suter, & T. P. Traas (Eds.), *Species sensitivity distributions in ecotoxicology* (p. 587). Lewis Publishers.

Teien, H. C., Garmo, A. C., Atland, A., & Salbu, B. (2008). Transformation of iron species in mixing zones and accumulation on fish gills. *Environmental Science & Technology*, *42*, 1780–1786.

U.S. Environmental Protection Agency. (1976). Quality criteria for water (p. 534).

U.S. Environmental Protection Agency. (1985). *Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses* (p. 98). Environmental Research Laboratory.

U.S. Environmental Protection Agency. (2018). *Final aquatic life ambient water quality criteria for aluminum 2018* (p. 329). Office of Water.

Van Sprang, P. A., Verdonck, F. A. M., Van Assche, F., Regoli, L., & De Schamphelaere, K. A. C. (2009). Environmental risk assessment of zinc in European freshwaters: A critical appraisal. *Science of the Total Environment*, *407*, 5373–5391.

van Straalen, N. M., & Van Leeuwen, C. J. (2002). European history of species sensitivity distributions. In L. Posthuma, G. W. Suter, & T. P. Traas (Eds.), *Species sensitivity distributions in ecotoxicology* (p. 587). Lewis Publishers.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*, 3–14.